УДК 811.161.1; 81-25; 81’322
DOI 10.17223/18137083/64/18

С. С. Земичева, **Е. В. Иванцова**

Томский государственный университет

Проект создания Томского диалектного корпуса в свете тенденций развития корпусной лингвистики [*]

Представлена концепция диалектного корпуса, репрезентирующего речь русских диалектоносителей Сибири. Показано, что проект исследователей Томского государственного университета отражает общие тенденции развития мировой и российской корпусной лингвистики, в то же время отличаясь рядом параметров. Новизна разрабатываемого корпуса определяется объектом представления (говоры обширной территории Среднего Приобья), репрезентативностью (архив 70-летнего экспедиционного обследования около 400 сел региона), лексикоцентрической и текстоцентрической ориентацией, структурой ресурса, характером подачи и разметки материалов устной речи. Обосновываются принципы создания Томского диалектного корпуса и сферы его использования.

Ключевые слова: корпусная лингвистика, Томский диалектный корпус, русские говоры Сибири.

Корпусная лингвистика как за рубежом, так и в России относится к числу наиболее актуальных сфер научного поиска. Корпусные разработки, как и словари, становятся не только источником данных, но и одним из эффективных методов лингвистического исследования [Perkuhn, et al., 2012, p. 19]. В настоящее время мировой перечень лингвистических корпусов весьма обширен, они базируются на разном материале и предполагают решение разных задач. В то же время можно выявить некоторые закономерности и лакуны в рассматриваемой области науки.

[*] Исследование выполнено при финансовой поддержке Российского научного фонда (проект № 16-18-02043).

Земичева Светлана Сергеевна – кандидат филологических наук, научный сотрудник лаборатории общей и сибирской лексикографии Томского государственного университета (просп. Ленина, 36, Томск, 634050, Россия; optysmith@gmail.com)

Иванцова Екатерина Вадимовна – доктор филологических наук, профессор кафедры русского языка Томского государственного университета (просп. Ленина, 36, Томск, 634050, Россия; ekivancova@yandex.ru)

ISSN 1813-7083. Сибирский филологический журнал. 2018. № 3

1. Обзор существующих диалектных корпусов и их место среди других корпусных ресурсов

В составе национальных корпусов преобладают письменные тексты: так, в Британском национальном корпусе (BNC) на долю устной речи приходится около 10 млн словоупотреблений, или 17,8 % от общего объема корпуса[1]. В Национальном корпусе русского языка (НКРЯ) объем устного корпуса также около 10 млн словоупотреблений, что составляет, однако, всего 2,8 % от его общего объема[2]. Из известных нам корпусов наиболее обширный материал устной речи включает корпус современного американского английского – 109 млн словоупотреблений, или 20 % всего корпуса[3]. Устная речь при этом понимается неоднозначно: для формирования и пополнения соответствующих подкорпусов используются прежде всего те тексты, которые уже представлены в расшифрованном и оцифрованном виде, в том числе записи теле- и радиопередач, стенограммы официальных мероприятий, переписка на интернет-форумах, фольклорные тексты, а также записи уроков, лекций, телефонных разговоров и т. п.

Создаются также корпуса, представляющие региолекты отдельных территорий. В качестве примера можно назвать банк «Голоса Юга», являющийся составной частью Американского национального корпуса[4], проект «Устная речь Финляндии: Разговорный язык в районе Хельсинки в 1972–1974 годах»[5], корпус разговорной речи Парижа[6] и др. В России на протяжении нескольких лет реализуется проект «Один речевой день», в рамках которого изучается речь жителей г. Санкт-Петербурга. По данным 2016 г. объем корпуса составлял более 1 200 часов звучания и около 1 млн словоупотреблений текстовых расшифровок [Русский язык повседневного общения, 2016, с. 14]. Создан также небольшой (около 40 минут звучания, 5 000 словоупотреблений) корпус «Рассказы сибиряков о жизни»[7]; существует проект Томского регионального корпуса [Резанова, 2015]; разрабатывается концепция звукового корпуса русской речи различных регионов России [Ерофеева и др., 2015].

Диалектные подкорпуса в большинстве известных европейских и американских корпусов отсутствуют. Лишь в некоторых из них, например в Чешском и Британском национальных корпусах, при репрезентации устной речи предусмотрена возможность поиска по территории, что позволяет изучать зональное варьирование языка.

Создание диалектных корпусных ресурсов, таким образом, представляет собой актуальную задачу. Необходимость их разработки связана, думается, с поисками истоков национального самосознания, возрастающей потребностью современного человека в самоидентичности в условиях технизации, стандартизации, широкого распространения массовой культуры, приводящих к обезличиванию индивида.

Корпусным исследованием диалектов занимаются лингвисты Германии, Испании, Португалии, Польши, Болгарии, Финляндии, Норвегии, Швеции, Грузии, Китая. Результатом их деятельности стало множество созданных баз данных и корпусов диалектной речи. Диалекты Британии были исследованы в этом аспекте одними из первых, работа велась параллельно в нескольких странах. Результаты реализации проекта британских ученых по исследованию английских диалек-

[1] http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=numbers
[2] http://www.ruscorpora.ru/corpora-stat.html
[3] http://corpus.byu.edu/coca/help/texts.asp
[4] http://newsouthvoices.uncc.edu/nsv
[5] http://www.ling.helsinki.fi/uhlcs/readme-all/README-uralic-lgs.html#C34
[6] http://cfpp2000.univ-paris3.fr/Corpus.html
[7] http://www.spokencorpora.ru/showcorpus.py?dir=01life

тов – «Survey of English Dialects» (SED) – представлены на сайте национальной библиотеки Великобритании в виде собрания аудиофайлов, снабженных коротким описанием[8]. Бóльшая часть материала собрана по вопросникам в 50–60 гг. XX в. Всего представлено 287 интервью из разных регионов продолжительностью около пяти минут каждое. Есть возможность выбрать определенный регион или год записи. Достаточно репрезентативен Хельсинкский корпус британских диалектов, который строится на записях 70–80-х гг., сделанных учеными из Финляндии. Было обследовано 92 населенных пункта в шести районах страны, опрошено 237 информантов, зафиксировано 846 149 словоупотреблений[9]. В Германии создан Фрайбургский корпус английских диалектов. Работа над ним ведется с 2000 г., заявленный объем корпуса – 2,3 млн словоупотреблений, однако материалы не представлены в свободном доступе из-за ограничений авторского права[10]. Существует также корпус письменных и устных шотландских текстов[11], где имеются возможности поиска по слову, доступа к полным текстам, прослушивания аудиозаписей.

На материале немецкого языка созданы банк данных разговорного немецкого языка, включающий диалектный подкорпус[12] и база данных баварских диалектов немецкого языка, объем которой оценивается создателями в диапазоне от 4 до 5 млн записей[13]. Проект «The Nordic Dialect Corpus»[14] содержит материалы диалектов нескольких скандинавских языков – норвежского, шведского, датского, фарерского, исландского.

Диалектные корпуса созданы также на материале других языков: испанского – «Corpus Oral y Sonoro del Español Rural»[15], португальского – «The Syntax-oriented Corpus of Portuguese Dialects»[16], болгарского – «Bulgarian Dialectology as Living Tradition»[17], польского – «Dialekty i gwary polskie. Kompendium internetowe»[18], грузинского[19]. В Китае, как указывают исследователи, наиболее активно изучается мандаринский диалект (Путунхуа), что связано с экстралингвистическими причинами, восприятием его как наиболее престижной разновидности китайского [Zu et al., 2002; Newman et al., 2008]. Современный мандаринский диалект китайского языка представлен в Ланкастерском корпусе, включающем письменные тексты[20], и корпусах устной речи: «Chinese Annotated Spontaneous Speech Corpus» (CASS), «Lancaster Los Angeles Spoken Chinese Corpus» (LLSCC); существует также корпус диалекта Вэньчжоу – «Wenzhou Spoken Corpus» (WSC)[21], разрабатывался проект мультидиалектного китайского корпуса [Zu et al., 2002].

Создано несколько диалектных корпусов русского языка. Лишь немногие из них включают материалы из разных регионов страны: диалектный подкорпус в составе НКРЯ[22], электронная база данных по русским говорам[23], акустическая

[8] http://sounds.bl.uk/Accents-and-dialects/Survey-of-English-dialects
[9] http://www.helsinki.fi/varieng/CoRD/corpora/Dialects/basic.html
[10] http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/
[11] http://www.scottishcorpus.ac.uk/advanced-search/
[12] http://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome
[13] http://www.baydat.uni-wuerzburg.de:8080/cocoon/baydat/projektinfo_BayDat
[14] http://www.tekstlab.uio.no/nota/scandiasyn/index.html
[15] http://www.lllf.uam.es/coser/index.php
[16] http://www.clul.ulisboa.pt/en/10-research/314-cordial-sin-corpus
[17] http://bulgariandialectology.org/
[18] http://www.dialektologia.uw.edu.pl/
[19] http://www.corpora.co/#/
[20] http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/
[21] http://ntuprojects.com/wenzhou/
[22] http://www.ruscorpora.ru/search-dialect.html
[23] http://www.ruslang.ru/agens.php?id=krylov_dialect

база данных «Русские регионы»[24]. Преобладают корпуса, созданные на материале конкретных регионов: корпус говоров р. Устья Архангельской области[25], Кубанский диалектный корпус [Трегубова, 2015], Саратовский диалектологический корпус [Крючкова, 2007], вологодский мультимедийный корпус «Жизненный круг» [Задумина, 2004], электронный текстовый корпус лингвокультуры Северного Приангарья[26].

Теоретические проблемы создания диалектных корпусов связаны в первую очередь со спецификой языковой системы местных говоров, имеющей значительные отличия от других форм национального языка. Создание диалектного электронного корпуса, как отмечает Т. Н. Москвина, сопряжено с целым рядом сложностей, среди которых «системные языковые отличия от литературного языка; исключительно устный характер диалектной коммуникации, как следствие – невозможность опереться на письменные источники; вариативность на всех уровнях, затрудняющая идентификацию единиц в корпусе», а также «собственно диалектная лексика, не поддающаяся простому переводу на литературный язык» [Москвина, 2014]. Затрудняет процессы формирования областных корпусов также собирание материалов для них в условиях языковой среды, к которой, как правило, не принадлежат диалектологи, трудоемкость экспедиционного сбора и обработки полученных данных, необходимость введения дополнительных параметров структуры и разметки корпуса, нерелевантных для кодифицированного языка, и мн. др.

В связи с обозначенными сложностями опережающими темпами развивается корпусная лингвистика на материале литературной письменной речи; диалектных корпусов в большом семействе электронных баз данных значительно меньше, чем основанных на литературном материале; значительно меньше и их объем; ждут решения многие теоретические проблемы, возникающие в связи с их созданием. Специфика материалов, имеющихся в том или ином региональном центре, накладывает отпечаток на каждую базу данных, несмотря на стремление к унификации корпусной продукции. Научные интересы лингвистической школы, в рамках которой создается новый электронный ресурс, также требуют выработки собственного подхода для решения данной задачи.

Разрабатываемая в Томском государственном университете концепция Томского диалектного корпуса (ТДК) вписывается в общий процесс развития корпусной лингвистики в целом и диалектных баз данных в частности, в то же время отличаясь от последних по ряду параметров.

2. Концепция Томского диалектного корпуса

Новизна ТДК определяется несколькими обстоятельствами.

2.1. Регион. Развивающаяся корпусная лингвистика в России опирается в основном на диалектные материалы европейской части страны. Это касается и сводных корпусов (НКРЯ, «Русские регионы»), где говоры восточнее Урала представлены единичными текстами, и локально ограниченных (Саратовская обл., Псковская обл., Кубань, Удмуртия, Вологодская обл., Архангельская обл.).

Существующие корпуса, созданные на материале сибирских говоров, пока весьма малы по объему. В частности, речь сибиряков отражена в уже упоминавшемся корпусе «Рассказы сибиряков о жизни»[27]. В 2017 г. создан электронный

[24] http://rureg.hs-bochum.de
[25] http://parasolcorpus.org/Pushkino/index.php
[26] http://angara.sfu-kras.ru/?page=dialect#
[27] http://www.spokencorpora.ru/showcorpus.py?dir=01life

текстовый корпус лингвокультуры Северного Приангарья (202 текста, 60 тыс. словоупотреблений)[28].

ТДК является, таким образом, одним из первых опытов создания диалектного корпуса, в котором репрезентированы данные сибирских говоров. Его разработка вписывается в мультидисциплинарный проект изучения уникального природного и социокультурного ареала Сибири в исследовательском центре «Транссибирский научный путь»[29].

В ТДК представлена речь русских старожилов на территории средней части среднеобского бассейна. Это обширный регион, охватывающий села по течению рек Томи, Оби и их притоков, в границах современного административного деления относящиеся к Томской и центральным районам Кемеровской обл. [Русские говоры..., 1984, с. 5]. Русское население закрепляется здесь с XVII в., после присоединения Западной Сибири к России – первоначально в нескольких острогах с приписанными к ним населенными пунктами, позднее распространяясь на близлежащие земли. Традиционная культура русского старожильческого населения Приобья, в том числе языковая, «представляет собой своеобразный феномен, сформировавшийся в особых природных условиях на основе тесного взаимодействия с автохтонными народами и потому значительно отличающийся от этнокультурного облика центральных районов России»[30]. Русские говоры Сибири являются вторичными, сложившимися в результате взаимного влияния речи старожилов и более поздних переселенцев – носителей русско-европейских материнских говоров с языком аборигенов края [Русские говоры..., 1984, с. 15].

2.2. Источники и репрезентативность корпуса. Одним из основополагающих принципов создания любого корпуса является его репрезентативность, которая «гарантирует типичность данных и обеспечивает полноту представления всего спектра языковых явлений» [Захаров, 2005, с. 3]. Сведения о репрезентативности диалектных корпусов, представленные в общедоступных источниках, не всегда дают возможность сопоставить материал по объему, так как в одних случаях указывается количество часов записи, в других – количество текстов, в третьих – количество словоупотреблений. Параметрами репрезентативности диалектного корпуса, кроме объема материала, являются также число информантов, количество обследованных населенных пунктов, продолжительность временно́го периода осуществления записей.

Объем диалектных корпусов варьируется весьма сильно. Так, корпус бесермянского диалекта удмуртского языка насчитывает около 60 тыс. словоупотреблений[31]. Наиболее обширные диалектные корпуса из известных нам[32] – база данных баварских диалектов немецкого языка (обследовано в общей сложности 1 613 баварских деревень, получено около 4 млн ответов на вопросники)[33], корпус шотландских текстов, насчитывающий более 4,5 млн словоупотреблений[34], и диалектный корпус скандинавских языков, содержащий около 2,8 млн слов из разговоров и интервью[35]. Объем около миллиона словоупотреблений можно, по-видимому, считать средним для диалектного корпуса. Так, Грузинский диалектный

[28] http://angara.sfu-kras.ru/?page=dialect#

[29] http://tssw.ru

[30] *Зенько А. П.* Русские старожилы Среднего Приобья: на стыке культур // Культурное наследие Югры: Электронная антология. URL: http://hmao.kaisa.ru/ object/1808928043?lc=ru

[31] http://beserman.ru/corpus/search/?interface_language=ru

[32] Если не брать в расчет web-корпуса, созданные на материалах, размещенных в Интернете.

[33] http://www.baydat.uni-wuerzburg.de:8080/cocoon/baydat/projektinfo_BayDat

[34] http://www.scottishcorpus.ac.uk/advanced-search/

[35] http://www.tekstlab.uio.no/nota/scandiasyn/index.html; см. также [Johannessen et al., 2012].

корпус насчитывает 1 871 459 слов, Эстонский диалектный корпус – 1 284 000 слов[36], в Ланкастерско-лос-анджелесский корпус разговорного китайского входит 1 002 151 слово[37]. Хельсинкский корпус британских диалектов включает 1 008 641 словоупотреблений[38].

Диалектный подкорпус НКРЯ, судя по статистике, пока не отличается ни достаточным объемом (197 текстов, или около 200 000 словоупотреблений)[39], ни пропорциональностью представления говоров разных территорий и типов. Складывается парадоксальная ситуация, при которой созданные корпуса отдельных территорий близки по объему к диалектному подкорпусу русского языка, цель которого – охватить территорию страны в целом. Так, корпус говоров р. Устья насчитывает более 800 000 словоупотреблений[40].

В отношении ТДК можно сказать, что корпус базируется на экспедиционных материалах 70-летнего изучения среднеобских говоров[41], обследовании около 400 сел региона, архивных записях (1 300 тетрадей, 200 часов звучания), что позволяет считать его достаточно репрезентативным в плане охвата материала. На данный момент в корпус входит более 600 текстов, около 700 тыс. словоупотреблений. Основная часть материалов находится в закрытом доступе, в свободное пользование предоставлена демонстрационная версия корпуса[42].

Вместе с тем в связи с экстралингвистическими причинами строгой сбалансированностью представления материалов различных временных срезов, групп говоров (нарымские, прикетские, приобские, притомские, причулымские) и говоров отдельных сел ТДК не отличается. Следует отметить также, что в течение многих лет основной целью диалектологов было полевое исследование только русских старожильческих говоров региона, носителями которых являются потомки первых поселенцев. Речь диалектоносителей более поздних волн переселения фиксировалась в меньшей степени. Записанные тексты в основном представляют собой «полуаутентичные», «провоцируемые» тексты с заданной собирателями темой коммуникации и вкраплениями спонтанной речи, типичные для условий полевого сбора материала.

2.3. Ориентация корпуса. Отражая этапы развития лингвистики в целом и диалектологии в частности, основная часть созданных диалектных корпусов ориентирована на представление системно-структурных особенностей местных говоров. Ряд диалектных корпусов (корпуса китайского языка, база данных баварских диалектов и др.) предназначен, прежде всего, для фонетических исследований. Основным видом разметки в большинстве случаев является морфологическая.

Ярким примером такого подхода к репрезентации местных говоров является диалектный подкорпус НКРЯ. Его принципы базируются на последовательном сравнении русских диалектов с литературным языком – прежде всего в области морфологии и лексики; с этой целью разработана система маркеров, выделяющих грамматические и лексические территориальные отличия от кодифицированной языковой подсистемы [Летучий, 2005, с. 215]. После недавней частичной коррек-

[36] http://www.murre.ut.ee/estonian-dialect-corpus/

[37] http://www.lancaster.ac.uk/fass/projects/corpus/LLSCC/

[38] http://www.helsinki.fi/varieng/CoRD/corpora/Dialects/

[39] http://www.ruscorpora.ru/corpora-stat.html

[40] http://parasolcorpus.org/Pushkino/stats.php

[41] Систематические полевые выезды для собирания диалектного материала стали осуществляться в Томском университете с 1946 г. [Томская диалектологическая школа, 2006, с. 16–20] и продолжаются по сей день. Недавно в распоряжение томских диалектологов поступили копии рукописных материалов экспедиций проф. А. Д. Григорьева, впервые осуществившего лингвистическое обследование этого региона в 1917–1922 г.

[42] http://losl.tsu.ru/?q=corpus/demo

тировки концепции этого ресурса появилась возможность обращения исследователя к полному тексту [Качинская, Сичинава, 2015].

Вместе с тем развитие науки о языке выдвигает перед областными корпусами новые задачи. Движение лингвистики в направлении от структурной к функциональной и когнитивной парадигмам вызывает необходимость изучения дискурсивных практик носителей языковой системы, исследования типов организации текста, отражения в них картины мира, мировосприятия и миропонимания homo loquens, выявления особенностей коммуникации в зависимости от социальной среды, условий общения и т. д. Активно анализируется метаязыковая рефлексия носителей языка, ставшая предметом перцептивной диалектологии [Anders et al., 2010; Александров, 2013].

Усиливается внимание к проблеме языка и культуры, оформляются как самостоятельные области знания лингвокультурология и этнолингвистика. На рубеже XX–XXI столетий формируется коммуникативная диалектология. В ней «вырабатывается новый подход к пониманию специфики диалекта, согласно которому своеобразие говора не сводится к его структурным особенностям в области фонетики, грамматики и лексики, а проявляется также в строении диалектных текстов, в соотношении различных жанров в составе диалектной коммуникации, в особых приемах раскрытия темы, в когнитивных особенностях диалектной речи, в особой картине мира, реализуемой в общении на диалекте» [Крючкова, 2007]. Все большее внимание (в том числе и в диалектной лексикографии) уделяется недифференциальному анализу местных говоров, общим принципом которого является изучение не только диалектных черт, но и общерусских элементов речи диалектоносителей, системных связей всех единиц лексикона.

Эти новые веяния нашли отражение и в сфере создания новых электронных ресурсов. Диалектные корпуса, существующие как в России, так и за рубежом, имеют несколько иную ориентацию по сравнению с корпусами литературных текстов. В болгарском, эстонском, скандинавском, шотландском, португальском диалектных корпусах предусмотрены как поиск по слову, так и просмотр целостных текстов, а также прослушивание аудио. Диалектные корпуса испанского и польского языков представляют собой, по сути, библиотеки текстов: поиск по слову в них невозможен, но представлены целостные тексты и аудиофайлы. В других случаях (Грузинский диалектный корпус и др.) возможен только поиск по слову, не предусмотрено обращение к целостным текстам. В целом же текстоцентрическую направленность и мультимодальность (доступ к звуковым файлам, интерактивным картам, фотографиям) можно считать типичной для зарубежных диалектных корпусов.

Создаваемый в России Саратовский диалектологический корпус ставит своей целью моделирование коммуникации в конкретных говорах, репрезентирующих специфику традиционной русской культуры сельского общения. Решение этой задачи осуществляется путем подачи текстов на широком культурном фоне, с привлечением исторических, географических, этнографических сведений, подробном комментировании упоминаемых в речи носителей говора событий, лиц, природных объектов, артефактов и т. п. [Крючкова, Гольдин, 2011]. Лингвокультурологическую направленность имеют также Электронный корпус диалектной культуры Кубани, отражающий тематически ориентированные фрагменты регионального дискурса («Обрядовая культура», «Традиционные верования», «Промысловая культура», «Бытовая культура» и др.) [Трегубова, 2015] и электронный текстовый корпус лингвокультуры Северного Приангарья[43].

Томский диалектный корпус также вписывается в новую лингвистическую проблематику. Он задуман с целью изучения своеобразия традиционной народно-

[43] http://angara.sfu-kras.ru/?page=dialect#

речевой культуры, репрезентированной в дискурсивной практике носителей сибирских старожильских говоров Среднего Приобья. Эта направленность обусловлена как общими процессами развития науки о языке, в том числе корпусной лингвистики и диалектологии, так и сферой интересов исследователей томской диалектологической школы. Ориентация на текст как единицу представления диалектного дискурса дает возможность изучать тематику общения на диалекте, систему речевых жанров, метаязыкового сознания диалектоносителей, своеобразие проявлений речевой культуры, роли фольклора в повседневной речи сельчан, влияния интенционального дискурса на бытовую личностно-ориентированную сферу общения и др.

Создаваемый текстоориентированный корпус одновременно можно охарактеризовать как лексикоориентированный. Несмотря на то, что в среднеобских говорах детально описаны все ярусы языковой системы, одним из центральных объектов анализа на протяжении всего периода их изучения является лексика. При этом от выявления собственно диалектных лексем и создания дифференциальных толковых словарей в 50–70-е гг. XX в. диалектологи перешли в 80–90-е гг. к описанию системных связей лексических единиц говора и составлению словарей полного типа, а в последние десятилетия – к лингвокультурологическому анализу диалектной концептосферы, реконструкции ментальных черт языковой личности диалектоносителя. Эти задачи также решаются с опорой прежде всего на лексические средства их выражения с учетом семантики, сочетаемости и контекста.

3. Представление материалов, структура ТДК и виды разметки

Своеобразие имеющегося архива, формировавшегося диалектологами в течение многих десятилетий, связано с последовательным отражением на разных этапах экспедиционной работы различных форм сохранения устной речи в полевых условиях: от ручной блокнотной фиксации (в том числе в транскрибированном виде) до регистрирования связных текстов диалектоносителей на магнитной ленте и цифровых носителях. Сложная задача их унифицирования решается через оцифровку всех сохранившихся аудиоматериалов экспедиций прошлых лет и переведение в электронный набор всех видов экспедиционных записей. В целях единообразной подачи разнородных первичных данных в качестве базового способа представления звучащей речи принята орфографическая запись с передачей отдельных региональных особенностей. При этом предусмотрен доступ к первоисточникам: просмотр сканированных рукописных текстов (для ранних записей) или прослушивание имеющихся аудиофайлов (для поздних).

В качестве базовой макроформы представления материала в корпусе избран текстовый файл, отражающий полный эпизод общения диалектоносителя с собирателем. Пользователям корпуса будут доступны как фрагменты текста, так и целостный файл. Текст представлен в традиционном для томской диалектологической школы орфографизированном виде, сохраняющем отличные от литературной нормы черты произношения и грамматики (долгие твердые шипящие, цоканье, стяженные формы глаголов и прилагательных и т. п.). Отсутствие транскрибированной расшифровки аудиозаписей компенсируется возможностью доступа к звуковым файлам. Отмечаются нераспознанные фрагменты звучащей речи, вопросы и реплики собирателей материала при диалогическом общении с информантами; даются комментарии диалектологов, способствующие пониманию ситуации и содержания текста. При наличии соответствующих материалов предполагается также дополнение текстовой части рисунками, фотографиями.

Некоторые электронные базы данных (в частности, корпус грузинских диалектов) используют для расширения материала иллюстрации из опубликованных областных словарей [Беридзе, Надараиа, 2011]; вологодский корпус также вклю-

чает тексты местной публицистики и беллетристики [Задумина, 2004]. Хотя среднеобский регион является одним из наиболее полно отраженных в диалектной лексикографии, такой способ пополнения ТДК не рассматривался: иллюстративные материалы словарей не отвечают принципу включения целостных, связных текстов. Вместе с тем идея связки корпус – словарь может быть реализована в другом виде. Планируется перевод опубликованных диалектных толковых словарей изучаемого региона в электронный формат, создание поисковой системы по этим словарям и ее привязка к текстовому корпусу. Это позволит в перспективе и решить задачу представления семантики областных слов в ТДК, и более эффективно использовать корпус для развития лексикографической базы (уточнение значения зафиксированных слов, пополнение иллюстративной части словарей, включение новых словарных статей). Таким образом, архитектоника корпуса, который на первом этапе разработки будет включать дешифрованные тексты, звуковые материалы и сканированные блокнотные записи, впоследствии дополнится лексикографическим разделом. Аналогичный подход представлен, например, в Болгарском диалектном корпусе, где имеется перевод на английский язык, и в диалектных корпусах, созданных на материале различных языков народов России, – например вепсского языка, где имеется перевод на русский.

Принципы разметки в ТДК имеют как достаточно стандартные черты, так и нововведения. Каждый вводимый в корпус текст подвергается трем типам разметки: паспортной, тематической и разметке по типу текста.

Паспортная разметка отражает экстралингвистические данные о времени, месте и характере записи, языковой личности информанта. Она включает дату сбора материала, населенный пункт, основные (ФИО, пол, год рождения) и дополнительные (образование, род занятий, места длительного проживания, информация о родителях и предках) сведения о диалектоносителе, архивный номер тетради.

Тематическая разметка в ТДК менее традиционна. Ее осуществление тесно связано с разработкой принципов тематического членения устной речи вообще и диалектной в том числе, представляющей собой сложную теоретическую задачу. В рамках корпусной лингвистики она еще только начинает решаться.

Существует точка зрения, что содержание включенных в корпус текстов не представляет интереса для лингвистов[44]. Однако представляется, что в свете новых задач коммуникативной диалектологии оно не менее важно, чем формальные параметры дискурса. Отмечается и значимость разнообразия тематики корпуса для семантических исследований [Москвина, 2014].

Наиболее простым способом представления отдельных тем диалектного дискурса является вычленение его фрагментов по принципу тематических блоков (как в лингвокультурологических кубанском и вологодском корпусах) или монотематического сборника (как, например, в электронной базе данных «Устные рассказы о Великой Отечественной войне»[45]); при этом отражение тем оказывается избирательным. Создатели НКРЯ опираются на общий для всех частей корпуса достаточно обобщенный список тем, исходя из тезиса о том, что в речи диалектоносителей «набор тем текстов мало отличается от литературного, но, естественно, гораздо более ограничен», а «диалектные тексты посвящены почти исключительно быту и обычаям» [Летучий, 2005, с. 230]. Это положение не может, на наш взгляд, рассматриваться как аксиома, а должно быть результатом анализа обширного материала народной речи. Кроме того, излишняя обобщенность выделения тем плохо соотносится с конкретностью мышления, характерной для диалектоносителей. Все включенные в национальный корпус диалектные тексты практически

[44] http://www.ruscorpora.ru/corpora-intro.html
[45] http://nocpskoviana.pskgu.ru/war.php

монотематичны, поскольку представляют собой сегменты полевых записей. Саратовские исследователи в основном следуют перечню тем национального корпуса с целью унификации данных при последующем сопоставлении. Вместе с тем они делают большой шаг вперед, исходя из реальности политематичной коммуникации, и указывают при разметке весь перечень затронутых в тексте тем в виде списка [Гольдин, Крючкова, 2006].

Текстовая разметка ТДК отличается как методикой, так и выделенным в конечном итоге составом тем. В качестве общих принципов разметки среднеобского диалектного дискурса по составу тем можно назвать следующие: вычленение тематики текста осуществлялось в направлении от частного к общему; иерархическое структурирование тем не превышало трех уровней (макротема – частная тема – коммуникативно значимая подтема); номинации тем по возможности соотносились с лексиконом рядового говорящего; при разметке использовалось «мягкое» членение, допускающее частичное наложение границ вычленяемых текстов. Состав тем также оказался иным, чем в диалектном подкорпусе НКРЯ и Саратовском диалектном корпусе: выделено 16 макротем («Работа», «Быт», «Еда», «Природа», «Происшествия» и др.) и 64 темы более частного порядка; в состав макротемы «Работа», например, входят темы «Обработка почвы», «Выращивание растений», «Заготовка кормов», «Выращивание животных», «Лесозаготовка», «Охота», «Ловля рыбы», «Шишкобой», «Сбор дикоросов», «Обработка льна», «Женские работы по дому», «Мужские работы по дому», «Прочие работы»; как высокочастотная в теме «Женские работы по дому» вынесена подтема «Рукоделие»). Отдельно маркировались атематические фрагменты, не отвечающие признакам связного текста, а также ситуативные включения, отражающие специфику устной коммуникации.

Кроме того, в ТДК введены виды разметки, которые пока не применяются в известных нам электронных базах данных. Получившая условное название «разметка по типам текста» отражает:

- метатекстовые фрагменты – «вербализованные суждения о языке как результат осознания языковой действительности» [Ростова, 2000, с. 55]. Высказывания такого рода дают представление об отношении носителей говоров к своей речи, восприятии речи окружающих, значении диалектных слов, их системных связях и функциональных характеристиках (мотивированное/немотивированное, новое/устаревшее, узуальное/неузуальное для говора, нейтральное/сниженное и т. п.);
- целенаправленную беседу с информантом по вопросникам. В данном случае маркируются фрагменты дискурса, наиболее далекие от естественной коммуникации диалектоносителей, но дающие лингвисту ценные сведения о семантике и употреблении лексических единиц, которые трудно выявить за короткие сроки в экспедиционных условиях. В комментарии отмечается характер вопросника: «Вопросник по теме “Растения”, “Обряды», “Рельеф”», «Вопросник для выявления мотивационных связей слов» и т. п.;
- диалог или полилог диалектоносителей. Это фрагменты дискурса, наиболее приближенные к естественной коммуникации жителей села;
- наиболее частотные речевые жанры бытовой коммуникации: автобиографический рассказ, рассказы о других лицах, рассказ о случае, воспоминание;
- встречающиеся в текстах разновидности фольклорных жанров: песни, частушки, пословицы и поговорки, приметы.

В настоящее время разработана концепция Томского диалектного корпуса, техническая документация и программное обеспечение к нему; создан электронный архив диалектных текстов, включающий сканированные ручные записи экспедиций 40–80-х гг. (более 1 000 единиц хранения), аудиотеку и видеотеку; пере-

веденные в компьютерный набор экспедиционные записи в объеме около 2 млн словоупотреблений; начат ввод текстов в корпус и их разметка.

Новый ресурс может быть использован при изучении русских народных говоров Сибири, обеспечивая доступ научной общественности к разнообразным материалам диалектологических экспедиций в Среднем Приобье, облегчая для исследователя трудоемкие задачи выборки данных и их системного анализа. Результаты работы над проектом внедряются в учебный процесс (практика по коммуникативистике для студентов-филологов, научно-исследовательская деятельность при обучении бакалавров, магистров и аспирантов), будут способствовать совершенствованию существующих диалектных словарей и созданию новой лексикографической продукции. Думается также, что Томский диалектный корпус внесет свой посильный вклад в исследование феномена народной речи во всем многообразии ее свойств.

Список литературы

Александров О. А. Диалектология восприятия: инновации в зарубежной лингвистике // Вестн. Иркут. гос. лингвистического ун-та. 2013. № 3(24). С. 52–58. URL: https://lib.mgpu.ru/materials/10/10912.pdf

Беридзе М. М., *Надарaиа Д. В.* Словарь как текстовый компонент корпуса (Корпус грузинских диалектов) // Тр. междунар. конф. «Корпусная лингвистика-2011», 27–29 июня 2011 г., С.-Петербург. СПб., 2011. С. 92–97. URL: https://events.spbu.ru/eventsContent/files/corpling/corpora2011/Beridze_92.pdf

Гольдин В. Е., *Крючкова О. Ю.* Тематическая разметка и тематический анализ диалектного текстового корпуса // Языковая личность – текст – дискурс: Теоретические и прикладные аспекты исследования: Материалы междунар. научн. конф.: В 2 ч. Ч. 1. Самара, 2006. С. 71–80.

Ерофеева Е. В., *Вардёй Б.*, *Краузе М.*, *Пост М.* Звуковой корпус региональной русской речи как инструмент изучения регионолектов и их оценки наивными носителями языка // Русский язык и литература в пространстве мировой культуры: Материалы XIII конгр. Междунар. ассоциации преподавателей рус. яз. и литературы (МАПРЯЛ), 13–20 сент. 2015 г., Гранада, Испания. СПб.: МАПРЯЛ; Гранада, 2015. Т. 2. С. 84–88.

Задумина П. Н. О некоторых особенностях создания мультимедийного корпуса региональных текстов // Молодые исследователи – регионам: Материалы междунар. науч. конф. Т. 3. Вологда, 2004. С. 194–196.

Захаров В. П. Корпусная лингвистика: Учеб.-методич. пособие. СПб., 2005. 48 с.

Качинская И. Б., *Сичинава Д. В.* Диалектный подкорпус сегодня // Тр. Ин-та рус. яз. им. В. В. Виноградова. Т. 6. М., 2015. С. 142–162.

Крючкова О. Ю. Электронный корпус русской диалектной речи и принципы его разметки // Изв. Саратов. ун-та. Новая сер. Филология. Журналистика. 2007. Т. 7, вып. 1. С. 30–34. URL: http://sarteorlingv.narod.ru/dialekt/elektr_korpus.html

Крючкова О. Ю., *Гольдин В. Е.* Корпус русской диалектной речи: концепция и параметры оценки // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегод. междунар. конф. «Диалог», 25–29 мая 2011 г., Бекасово. Вып. 10(17). М., 2011. С. 359–367. URL: http://www.dialog-21.ru/media/1437/36.pdf

Летучий А. Б. Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005. С. 215–233. URL: http://ruscorpora.ru/sbornik2005/13letuchy.pdf

Москвина Т. Н. Методы и подходы корпусной лингвистики в исследованиях семантики диалектной лексики // Современные проблемы науки и образования.

2014. № 6. URL: http://www.science-education.ru/ru/article/view?id=15784 (дата обращения 10.05.2017).

Резанова З. И. Лингвистический корпус «Томский региональный текст»: типологически релевантные параметры сбалансированности и репрезентативности // Вестн. Том. гос. ун-та. Филология. 2015. № 1(33). С. 38–50.

Ростова А. Н. Метатекст как форма экспликации метаязыкового сознания. Томск: Изд-во Том. ун-та, 2000. 193 с.

Русские говоры Среднего Приобья / Под ред. В. В. Палагиной. Ч. 1. Томск: Изд-во Том. ун-та, 1984. 208 с.

Русский язык повседневного общения: особенности функционирования в разных социальных группах / Отв. ред. Н. В. Богданова-Бегларян. СПб.: Лайка, 2016. 244 с.

Томская диалектологическая школа: Историографический очерк / Под ред. О. И. Блиновой. Томск: Изд-во Том. ун-та, 2006. 392 с.

Трегубова Е. Н. Многоуровневая тематическая разметка как инструмент этнолингвистической репрезентации диалектного дискурса в электронном текстовом корпусе // Вестн. Том. гос. ун-та. Филология. 2015. № 1(33). С. 66–77.

Anders C. A., *Hundt M.*, *Lasch A.* Perceptual Dialectology. Neue Wege der Dialectologie. Berlin: Degruyter, 2010. 449 p.

Johannessen J. B., *Priestley J.*, *Hagen K.*, *Nøklestad A.*, *Lynum A.* The Nordic dialect corpus // Proc. of the Eighth Intern. Conf. on Language resources and Evaluation. 2012. P. 3387–3392. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/773_Paper.pdf

Newman J., Lin J., *Butler T.*, *Zhang E.* The Wenzhou spoken corpus // Corpora. 2008. Vol. 2, iss. 1. P. 97–109. URL: http://dx.doi.org/10.3366/cor.2007.2.1.97

Perkuhn R., *Keibel H.*, *Kupietz M.* Korpuslinguistik. Paderborn: Wilhelm Fink Verl., 2012. 144 p.

Zu Y., *Chen Y.*, *Zhang Y.*, *Zhou L.*, *Shen M.*, *Huang J.* A Super phonetic system and multi-dialect Chinese speech corpus for speech recognition // Proc. of Intern. Conf. on Spoken Language Processing. 2002. URL: http://www.colips.org/conferences/iscslp2006/anthology/2002/Papers/048.PDF

S. S. Zemicheva[1], **E.V. Ivantsova**[2]

Tomsk State University, Tomsk, Russian Federation
[1] *optysmith@gmail.com,* [2] *ekivancova@yandex.ru*

The project of Tomsk dialect corpus in keeping with trends of corpus linguistics development

The concept of the dialect corpus representing the Russian dialect speech of the Middle Ob region is proposed. The authors demonstrate that the project of Tomsk dialect corpus corresponds to the key trends of modern corpus linguistics: the involvement of oral speech materials; attention to the regional variation of the language; the study of dialect as part of the traditional culture; multimodality. The novelty of the resource is determined by the material – it is one of the few corpuses that include the speech of residents of the vast Siberian region: the archive includes the results of a 70-year expedition survey of about 400 villages – and lexicocentric and textocentric orientation: the possibility of access to full texts is fundamentally important. The problem of representativeness and balance of the dialect corpus which has not been studied in the scientific literature is considered. Today, Tomsk dialect corpus includes approximately 700 000 words, allowing it to be considered as a fairly representative collection of dialect texts. At the same time, the special characteristics of the material result in the corpus being not strictly balanced. The texts are

presented in spelling with some phonetical features of the dialect. The structure of the new electronic resource involves 3 types of markup: passport, thematic and type of text. Passport meta-markup includes extra-linguistic data about the texts: the place of recording, the date, the information about the informant (sex, age, place of birth, level of education, occupation). Thematic meta-markup is made by means of an inductive analysis of the discursive practices of old-timers. The list of topics is hierarchical, with each topic being three levels deep maximum. The principle of «soft» markup is used, with the possibility of simultaneously assigning several themes to the one text fragment. At the first level of the hierarchy, 16 macro-themes are marked (Work, Food, Nature, etc.), on the second – 64 topics. Firstly, the markup by type of text at this stage includes the degree of the spontaneity of speech events and, secondly, the most frequent speech genres. The prospects for using the resource are the study of Middle Ob dialects in linguocultural, genre, communicative, cognitive, linguopersonological and other aspects; the creation of new dialect dictionaries; the investigation of traditional culture and folklore, customs and rituals, history of the region.

Keywords: corpus linguistics, Tomsk dialect corpus, Russian dialects of Siberia.

DOI 10.17223/18137083/64/18

References

Aleksandrov O. A. Dialektologiya vospriyatiya: innovatsii v zarubezhnoy lingvistike [Dialectology of perception: innovations in foreign linguistics]. *ISLU Philological Review*. 2013, no. 3(24), pp. 52–58.

Anders C. A., Hundt M., Lasch A. *Perceptual Dialectology. Neue Wege der Dialec-tologie*. Berlin, Degruyter, 2010, 449 p.

Beridze M. M., Nadaraia D. V. Slovar' kak tekstovyy komponent korpusa (Korpus gruzinskikh dialektov) [Dictionary as the text component of the corpus (corpus of Georgian dialects)]. In: *Tr. mezhdunar. konf. "Korpusnaya lingvistika-2011", 27–29 iyunya 2011 g. S.-Peterburg* [Proceedings of the international conference "Corpus linguistics-2011" (June 27–29, 2011, St. Petersburg)]. St. Petersburg, 2011, pp. 92–97. URL: https://events.spbu.ru/eventsContent/files/corpling/corpora2011/Beridze_92.pdf

Erofeyeva E. V., Vardëy B., Krauze M., Post M. Zvukovoy korpus regional'noy russkoy rechi kak instrument izucheniya regiolektov i ikh otsenki naivnymi nositelyami yazyka [Sound corpus of the Russian regional speech as a tool for study regiolects and their evaluation by naive speakers]. In: *Russkiy yazyk i literatura v prostranstve mirovoy kul'tury: Materialy XIII kongr. Mezhdunar. assotsiatsii prepodavateley rus. yaz. i litera-tury (MAPRYAL), 13–20 sent. 2015 g., Granada, Ispaniya* [Russian language and literature in the space of world culture: Proceedings of the 13th congress of MAPRYAL Sept. 13–20, 2015, Granada, Spain]. St. Petersburg, MAPRYAL, Granada, 2015, vol. 2, pp. 84–88.

Gol'din V. E., Kryuchkova O. Yu. Tematicheskaya razmetka i tematicheskiy analiz dialectnogo tekstovogo korpusa [Theme markup and thematic analysis of the dialect text corpus]. In: *Yazykovaya lichnost' – tekst – diskurs: Teoreticheskiye i prikladnyye aspekty issledovaniya: Materialy mezhdunar. nauchn. konf.: V 2 ch. Ch. 1* [Linguistic personality – text – discourse: theoretical and applied aspects of research: proceedings of the intern. sci. conf.: in 2 pts. Pt 1]. Samara, 2006, pp. 71–80.

Johannessen J. B., Priestley J., Hagen K., Nøklestad A., Lynum A. The Nordic dialect corpus. In: *Proc. of the Eighth Intern. Conf. on Language resources and Evaluation.* 2012, pp. 3387–3392. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/ 773_Paper.pdf

Kachinskaya I. B., Sichinava D. V. Dialektnyy podkorpus segodnya [Dialect subcorpus today]. *Proceedings of the V.V. Vinogradov Russian Language Institute.* 2015, vol. 6, pp. 142–163.

Kryuchkova O. Yu., Gol'din V. E. Korpus russkoy dialektnoy rechi: kontseptsiya i parametry otsenki [The Corpus of Russian dialect speech: the concept and parameters of evaluation]. In: *Komp'yuternaya lingvistika i intellektual'nyye tekhnologii: Po materialam ezhegod. mezhdunar. konf. "Dialog", 25–29 maya 2011 g., Bekasovo. Vyp. 10(17)* [Computer linguistics and intellectual technologies: Based on materials of annual intern. conf. "Dialogue", May 25–29, 2011, Bekasovo. Iss. 10(17)]. Moscow, 2011, pp. 359–367. URL: http://www.dialog-21.ru/media/1437/36.pdf

Kryuchkova O. Yu. Elektronnyy korpus russkoy dialektnoy rechi i printsipy ego razmetki [Electronic corpus of Russian dialect speech and the principles of its markup]. *Izvestiya of Sara-*

tov University. New Series. Series: Philology. Journalism. 2007, vol. 7, iss. 1, pp. 30–34. URL: http://sarteorlingv.narod.ru/dialekt/elektr_korpus.html

Letuchiy A. B. Korpus dialektnykh tekstov: zadachi i problemy [Corpus of dialect texts: tasks and problems]. In: *Natsional'nyy korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [The National Corpus of the Russian language: 2003–2005. Results and prospects]. Moscow, 2005, pp. 215–233. URL: http://ruscorpora.ru/sbornik2005/13letuchy.pdf

Moskvina T. N. Metody i podkhody korpusnoy lingvistiki v issledovaniyakh semantiki dialektnoy leksiki [Methods and approaches of corpus linguistics in studies of the semantics of dialect vocabulary]. *Sovremennyye problemy nauki i obrazovaniya.* 2014, no. 6. URL: http://www.science-education.ru/ru/article/view?id=15784 (accessed 10.05.2017).

Newman J., Lin J., Butler T., Zhang E. The Wenzhou spoken corpus. In: *Corpora.* 2008, vol. 2, iss. 1, pp. 97–109. URL: http://dx.doi.org/10.3366/cor.2007.2.1.97

Perkuhn R., Keibel H., Kupietz M. *Korpuslinguistik. Paderborn*: *Wilhelm Fink Verl.,* 2012, 144 p.

Rezanova Z. I. Lingvisticheskiy korpus "Tomskiy regional'nyy tekst": tipologicheski relevantnyye parametry sbalansirovannosti i reprezentativnosti [Linguistic corpus "Tomsk regional text": typologically relevant parameters of balance and representativeness]. *Tomsk State University Journal of Philology.* 2015, no. 1(33), pp. 38–50.

Rostova A. N. *Metatekst kak forma eksplikatsii metayazykovogo soznaniya* [Metatext as a form of explication of metalanguage consciousness]. Tomsk, TSU, 2000, 193 p.

Russkiye govory Srednego Priob'ya. Ch. 1 [Russian dialects of the Middle Ob region. Pt 1]. V. V. Palagina (Ed.). Tomsk, TSU, 1984, 208 p.

Russkiy yazyk povsednevnogo obshcheniya: osobennosti funktsionirovaniya v raz-nykh sotsial'nykh gruppakh [Russian language of everyday communication: features of functioning in different social groups]. N. V. Bogdanova-Beglaryan (Ed.). St Petersburg, Layka, 2016, 244 p.

Tomskaya dialektologicheskaya shkola: Istoriograficheskiy ocherk [Tomsk school of dialectology: A historiographical sketch]. O. I. Blinova (Ed.). Tomsk, TSU, 2006, 392 p.

Tregubova E. N. Mnogourovnevaya tematicheskaya razmetka kak instrument etnolingvisticheskoy reprezentatsii dialektnogo diskursa v elektronnom tekstovom korpuse [Multilevel thematic marking as an ethnolinguistic tool of dialectal discourse representation in digital text corpora]. *Tomsk State University Journal of Philology.* 2015, no. 1(33), pp. 66–77.

Zadumina P. N. O nekotorykh osobennostyakh sozdaniya mul'timediynogo korpusa regional'nykh tekstov [On some features of creating a multimedia corpus of regional texts]. In: *Molodyye issledovateli – regionam: Materialy mezh-dunar. nauch. konf. T. 3.* [Young researchers to regions. Proceedings of the intern. sci. conf. Vol. 3]. Vologda, 2004, pp. 194–196.

Zakharov V. P. *Korpusnaya lingvistika: Ucheb.-metodich. posobiye* [Corpus linguistics: Educational and methodical manual]. St. Petersburg, 2005, 48 p.

Zu Y., Chen Y., Zhang Y., Zhou L., Shen M., Huang J. A Super phonetic system and multidialect Chinese speech corpus for speech recognition. In: *Proc. of Intern. Conf. on Spoken Language Processing*. 2002. URL: http://www.colips.org/conferences/iscslp2006/anthology/2002/Papers/048.PDF